AMaPO: Adaptive Margin-attached Preference Optimization for Language Model Alignment

Ruibo Deng[1,2], **Duanyu Feng**[1,2], **Wenqiang Lei**[1,2*]

[1]Sichuan University, Chengdu, China
[2]Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, China
dengruibo@stu.scu.edu.cn, fengduanyuscu@stu.scu.edu.cn, wenqianglei@scu.edu.cn

Abstract

Offline preference optimization offers a simpler and more stable alternative to RLHF for aligning language models. However, their effectiveness is critically dependent on ranking accuracy, a metric where further gains are highly impactful. This limitation arises from a fundamental problem that we identify and formalize as the Overfitting-Underfitting Dilemma: current margin designs cause models to apply excessive, wasteful gradients to correctly ranked samples (overfitting) while providing insufficient corrective signals for misranked ones (underfitting). To resolve this dilemma, we propose **Adaptive Margin-attached Preference Optimization (AMaPO)**, a simple yet principled algorithm. AMaPO employs an instance-wise adaptive margin, refined by Z-normalization and exponential scaling, which dynamically reallocates learning effort by amplifying gradients for misranked samples and suppressing them for correct ones. Extensive experiments on widely used benchmarks demonstrate that AMaPO not only achieves better ranking accuracy and superior downstream alignment performance, but targeted analysis also confirms that it successfully mitigates the core overfitting and underfitting issues.

Code — https://github.com/Shiroha-Offical/AMaPO
Extended version — https://arxiv.org/pdf/2511.09385

1 Introduction

Aligning Large Language Models (LLMs) with human preferences, ensuring they are helpful, honest, and harmless, also known as preference learning, is crucial for LLM research (Bai et al. 2022; Ouyang et al. 2022). A standard approach for this task is Reinforcement Learning from Human Feedback (RLHF), a powerful online algorithm that first trains a reward model on human preferences and then uses reinforcement learning to optimize the LLMs (Ouyang et al. 2022; Schulman et al. 2017). Despite its demonstrated success, RLHF is notoriously complex and prone to training instabilities, presenting significant practical challenges (Engstrom et al. 2020). To circumvent these issues, a new family of offline alignment algorithms has emerged, led by methods like DPO (Rafailov et al. 2023), IPO (Azar et al. 2024), KTO (Ethayarajh et al. 2024) and FocalPO (Liu et al. 2025). These approaches offer superior simplicity and stability by directly optimizing the LLMs to favor preferred responses over dispreferred ones, thereby bypassing the need for an explicit reward model and the complexities of RL. At its core, the efficacy of this entire paradigm is the ranking accuracy of an implicit reward model, which dictates how well the optimization process can distinguish between desirable and undesirable outputs.

Therefore, many subsequent DPO-style methods have been used to improve ranking accuracy, either explicitly prioritizing ranking (Chen et al. 2024) or indirectly fostering this capability through other modifications. A primary direction of these methods is to manipulate the implicit reward function, often by incorporating a reward margin to better separate preferred and dispreferred responses. Innovations in this direction include introducing fixed or dynamic margins (Meng, Xia, and Chen 2024; Wu et al. 2024) or updating the reference model to create more favorable optimization landscapes (Gorbatovski et al. 2024). Although these methods report improved performance, their contributions are typically validated empirically, without a unified theoretical perspective to explain how these methods dynamically affect ranking accuracy during training and to compare with each others. Separately, emerging work has analyzed the gradient dynamics of preference learning algorithms (Feng et al. 2024; Yan et al. 2024; Yuan et al. 2024; Ma et al. 2025). However, these studies often fail to explicitly connect their findings on gradient properties, such as magnitude or conflict, with the resulting changes in ranking accuracy. Therefore, we argue that a crucial analytical bridge is missing: a framework that not only examines the gradient dynamics of these advanced DPO variants but also directly links them to the dynamic evolution of ranking accuracy.

To bridge this analytical gap, we introduce a unified framework that analyzes DPO-style methods through the lens of the reward margin. First, we reformulate existing algorithms, including DPO and its variants, into a generalized objective function based on different margin designs. This unification establishes a direct, formal connection between the algorithmic structure of each method and its capacity to optimize ranking accuracy. Within this framework, we then employ gradient dynamics analysis to investigate how these distinct margin formulations shape the learning trajectory of

*Corresponding author

ranking accuracy. Our analysis reveals a fundamental flaw in the standard DPO mechanism: for samples that are already correctly ranked, its margin design leads to *overfitting* by assigning excessively large gradients, wasting capacity on these easy samples. Conversely, for incorrectly ranked samples where learning is most critical, the margin often results in *underfitting* by providing insufficient gradients, thereby hindering the correction of errors. We also find that while some of subsequent DPO-style methods attempt to address this, they fail to fully resolve this core tension.

Motivated by our analysis, we introduce AMaPO, an algorithm designed to resolve this overfitting-underfitting dilemma. AMaPO's core strategy is an instance-wise adaptive margin, refined by Z-normalization and exponential scaling. By dynamically assigning large corrective margins to underfit samples and a zero margin to correctly ranked ones, it effectively reallocates learning effort to where it is most needed. Our extensive experiments on diverse benchmarks demonstrate that AMaPO not only achieves state-of-the-art ranking accuracy and superior downstream alignment performance, but targeted analysis also confirms that it successfully mitigates the core overfitting and underfitting issues, validating our approach.

The main contributions of this paper are: (i) We introduce a unified margin-based framework to analyze gradient dynamics of DPO-style algorithms, which reveals an overfitting-underfitting dilemma: existing methods expend excessive gradient on already-learned preferences (overfitting) while providing insufficient signal to correct misranked ones (underfitting). (ii) We propose Adaptive Margin-attached Preference Optimization (AMaPO), a principled algorithm directly resolves this dilemma. AMaPO employs an instance-wise adaptive margin and refined by Z-normalization and exponential scaling, that reallocates learning effort by amplifying gradients for misranked samples and suppressing them for correctly ranked ones. (iii) Extensive experiments demonstrate AMaPO achieves state-of-the-art ranking accuracy, solving the overfitting-underfitting dilemma and leading to superior downstream alignment performance like instruction-following and complex reasoning.

2 Related Works

Offline Preference Learning Algorithms Offline Preference Learning Algorithms, initiated by Direct Preference Optimization (DPO) (Rafailov et al. 2023), directly optimize the LLM on preference data, obviating practical challenges such as the need for a separate reward model and training instability. Optimized DPO variants generally fall into two categories: reformulating the core loss function (Azar et al. 2024; Ethayarajh et al. 2024), or incorporating an explicit reward margin to separate preference pairs robustly (Boser, Guyon, and Vapnik 1992; Turner and Firth 2012). This margin-based approach has been realized through employing fixed or instance-dependent margins to improve learning stability and handle potential data noise (Meng, Xia, and Chen 2024; Kim et al. 2024; Amini, Vieira, and Cotterell 2024). Others achieve a form of dynamic regularization by progressively updating the reference policy itself, which effectively creates an adaptive margin throughout the training process (Gorbatovski et al. 2024; Wu et al. 2024). Despite these diverse, the central mechanism common to all these DPO-style algorithms is the implicit reward function. This makes their ability to learn to classify preferences correctly, known as the ranking accuracy (Chen et al. 2024), the paramount determinant of their success. However, the field currently lacks a unified framework to analyze how these different methods impact the evolution of ranking accuracy, leaving a critical gap in our understanding.

Preference Learning Theory Concurrent with the development of preference learning algorithms, a body of theoretical work has emerged to analyze their underlying mechanisms. These inquiries generally follow two primary perspectives: gradient dynamics and divergence analysis. From the gradient perspective, for example, (Feng et al. 2024) reveals the limited learning capacity of DPO through field theory and (Yuan et al. 2024) identify issues such as gradient entanglement in margin-based methods. From divergence perspective, it examines how DPO optimizes the model distribution π_θ towards the distribution of chosen response. While DPO implicitly optimizes a forward KL-divergence, recent work identifies promoting mode-seeking behavior by optimizing a reverse KL-divergence or Total Variation distance as a crucial property for preference alignment(Tajwar et al. 2024; Xiao et al. 2024, 2025). While these analyses are foundational, they typically examine each algorithm as a distinct entity. In this paper, we also want to build upon the tradition of gradient analysis but introduce a novel, unifying perspective. Based on the unification, we further explicitly connect the specific gradient properties of each method to their dynamic impact on ranking accuracy.

3 Preliminaries

This section establishes the technical foundations for our analysis. First, we detail DPO (Rafailov et al. 2023), the seminal algorithm for offline preference learning that serves as the basis for the methods we investigate. Second, we formalize the concept of the reward margin, which is used to construct a unified representation of various DPO-style algorithms. Finally, we define ranking accuracy, the primary metric used throughout our work to evaluate the efficacy of these alignment methods.

3.1 Direct Preference Optimization

Preference Data and the Bradley-Terry Model. Offline preference optimization relies on a static dataset of preferences, $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$. Each entry consists of a prompt $x^{(i)}$ and a pair of responses, $(y_w^{(i)}, y_l^{(i)})$, where y_w is preferred over y_l, denoted as $y_w \succ y_l$. This preference is typically provided by a human annotator or a powerful reward model. A standard assumption in the field is that these preferences are drawn from a distribution that follows the Bradley-Terry model (Bradley and Terry 1952), which models the probability of preferring y_w over y_l as a logistic function of the difference in their latent reward values: $p(y_w \succ y_l|x) = \sigma(r(x, y_w) - r(x, y_l))$.

The Direct Preference Optimization. Instead of explicitly training a reward model $r(x, y)$, DPO (Rafailov et al. 2023) derives an analytical mapping from the optimal policy to the reward function. This allows DPO to use the log-likelihood of the policy to implicitly represent the reward function via a closed-form expression with the optimal policy:

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\big[\log\sigma\big(\beta\log\tfrac{\pi_\theta(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \beta\log\tfrac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\big)\big], \quad (1)$$

where π_{ref} is a fixed reference model, and β is a temperature parameter that scales the reward.

3.2 A Unified Margin-based Framework

Recent theoretical works have sought to unify DPO-style algorithms under a single mathematical representation (Yuan et al. 2024; Liu, Liu, and Cohan 2024). A common approach is to formulate a general loss function that captures variations in reward shaping and regularization. However, with the increasing prominence of methods that explicitly engineer a reward margin (Zhao et al. 2024; Wu et al. 2024), these existing frameworks can represent such methods only indirectly, obscuring the margin's central role.

To create a more direct analytical tool that reflects this research trend, we extend the prior work (Yuan et al. 2024) by proposing a refined unified objective with the margin, γ:

$$\mathcal{L}_{\mathrm{unified}}(\theta) = -\left(m(h_w(\log\pi_w) - h_l(\log\pi_l) - \gamma) + \Lambda\left(\log\pi_w\right)\right). \quad (2)$$

Here, $h_w(x)$ and $h_l(x)$ represent transformation or adjustment functions applied to the log-probabilities of the preferred and dispreferred responses, respectively. π_w and π_l represent $\pi_\theta\left(y_w|x\right)$ and $\pi_\theta\left(y_l|x\right)$ for simplicity, respectively. The term $m(x)$ is a scoring function that computes the difference between these transformed values with the margin γ. Finally, $\Lambda\left(\log\pi_w\right)$ is an optional auxiliary term that some models incorporate to encourage additional learning specifically on the preferred response. This formulation allows us to directly compare algorithms based on their margin design. For instance:

DPO: The original DPO algorithm fits this framework with $h_w(x) = h_l(x) = \beta x$, $m(x) = \log\sigma(x)$ and a fixed margin of $\gamma = \beta(\log\pi_{\mathrm{ref}}(y_w|x) - \log\pi_{\mathrm{ref}}(y_l|x))$.

SimPO: Simple Preference Optimization (SimPO) (Meng, Xia, and Chen 2024) is a highly effective DPO variant, shown as

$$\mathcal{L}_{\mathrm{SimPO}}(\pi_\theta; \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\big[\log\sigma\big(\tfrac{\beta}{|y_w|}\log\pi_\theta\left(y_w|x\right) - \tfrac{\beta}{|y_l|}\log\pi_\theta\left(y_l|x\right) - C\big)\big]. \quad (3)$$

It fits our framework by defining $h_w(x) = \frac{\beta}{|y_w|}x$, $h_l(x) = \frac{\beta}{|y_l|}x$, $m(x) = \log\sigma(x)$ and a tunable margin $\gamma = C > 0$.

This margin-focused unification serves as the cornerstone of our gradient analysis. [1]

[1]For completeness, we formulate other common DPO-style methods within this framework in Extended version.

3.3 Ranking Accuracy

While the ultimate goal of alignment is to produce generations that humans prefer, which is often evaluated by win-rate against a baseline model (Zheng et al. 2023b), this metric is costly and ill-suited for tracking training dynamics. Therefore, a more direct and practical metric is crucial for analyzing the learning process. Ranking accuracy (Liu et al. 2024; Chen et al. 2024) directly measures whether the policy (training) model π_θ itself assigns a higher likelihood to the preferred response over the dispreferred one for a given prompt. Formally, it is defined as:

$$\mathcal{R}(\pi_\theta; \mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{(x,y_w,y_l)\in\mathcal{D}}\mathbb{I}\left[\pi_\theta(y_w|x) > \pi_\theta(y_l|x)\right], \quad (4)$$

where $\mathbb{I}[\cdot]$ is the indicator function. A higher ranking accuracy indicates that the policy's preference ordering aligns with the ground-truth data, and it serves as a predictor of the downstream performance within a certain KL divergence range (Chen et al. 2024). Furthermore, for our theoretical analysis, we utilize the instance-wise ranking accuracy derived from our unified framework (Eq. 2), defined as $h_w(\log\pi_w) - h_l(\log\pi_l)$. For the vast majority of DPO-style methods where the transformations h_w and h_l are linear or monotonic, the sign of this score is equivalent to the standard ranking accuracy (i.e., $h_w(\log\pi_w) - h_l(\log\pi_l) > 0 \iff \pi_\theta(y_w|x) > \pi_\theta(y_l|x)$). Throughout this paper, we use this ranking accuracy as the primary lens for analysis.

4 Methods

In this section, we leverage the unified margin-based framework established in the previous section to conduct a detailed theoretical analysis and construct our related method. Our primary goal is to understand how the margin design in DPO and its variants affects the learning dynamics of ranking accuracy. Through this gradient-level investigation, we diagnose the key limitations of existing approaches and, from these findings, distill a set of desiderata for a more effective margin function. Building on these principles, we then introduce **Adaptive Margin-attached Preference Optimization (AMaPO)**, a novel preference optimization algorithm that assigns an adaptive margin specifically designed to manipulate gradient magnitude, thereby improving both ranking accuracy and overall alignment performance.

4.1 Gradient Dynamics and the Ideal Margin

Based on the unified objective in Eq. 2, we can derive a general gradient formulation for DPO-style algorithms. The gradient with respect to the model parameters $\boldsymbol{\theta}$ is:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{unified}}\left(\theta\right) = -m'(h_w(\log\pi_w) - h_l(\log\pi_l) - \gamma)\left(h'_w(\log\pi_w) - h'_l(\log\pi_l)\right) - \Lambda'\left(\log\pi_w\right). \quad (5)$$

Specifically, we define $d_{\boldsymbol{\theta}} = m'(h_w(\log\pi_w) - h_l(\log\pi_l) - \gamma)$, which comprises log-probabilities of responses $(h_w(\log\pi_w),\ h_l(\log\pi_l))$ and margin γ and significantly affects the gradient magnitude[2].

[2]For simplicity, we omit the contribution of $\Lambda'(\log\pi_w)$ to the gradient magnitude, as $\Lambda'(\log\pi_w)$ is typically a constant and not related to γ.

Case	$r_{\pi_\theta}(x, y_w, y_l), \gamma$	$\lvert d_{\boldsymbol{\theta}} \rvert$	Influence
1	$r_{\pi_\theta}(x, y_w, y_l) \geq \gamma$	$\downarrow$	mild update on **correctly ranked** samples, **ideal**.
2	$r_{\pi_\theta}(x, y_w, y_l) < \gamma$	$\uparrow$	aggressive update on **correctly ranked** samples, **overfitting**.
3	$r_{\pi_\theta}(x, y_w, y_l) \geq \gamma$	$\downarrow$	mild update on **incorrectly ranked** samples, **underfitting**.
4	$r_{\pi_\theta}(x, y_w, y_l) < \gamma$	$\uparrow$	aggressive update on **incorrectly ranked** samples, **ideal**.

Table 1: All four possible cases for training dynamics of the unified framework in Eq. 2, where $\uparrow$ and $\downarrow$ indicate increase and decrease, $r_{\pi_\theta}(x, y_w, y_l) = h_w(\log \pi_w) - h_l(\log \pi_l)$. The **correctly ranked** means $r_{\pi_\theta}(x, y_w, y_l) > 0$, the **incorrectly ranked** means $r_{\pi_\theta}(x, y_w, y_l) \leq 0$. **Case 1** and **Case 4** are the ideal behavior which properly control the learning rate.

This insight reveals a crucial mechanism: for a given state of the policy π_θ, the **margin γ is the primary lever to control the learning rate**. Since the gradient magnitude $d_{\boldsymbol{\theta}}$ is governed by

$$d_{\boldsymbol{\theta}} = m'(h_w(\log \pi_w) - h_l(\log \pi_l) - \gamma) \tag{6}$$

where m' is usually a monotonically decreasing function and $m'(x) > 0$ (as $m(x) = \log \sigma(x)$), the choice of γ directly determines whether the learning signal is amplified or suppressed. A poorly designed margin leads directly to what we term the Overfitting-Underfitting Dilemma.

Definition 1 (The Overfitting-Underfitting Dilemma) *An algorithm is susceptible to this dilemma when its margin γ causes inefficient optimization:*

- ***Overfitting:*** *A large margin produces wasteful gradients for correctly ranked samples, forcing the model to over-optimize on these samples.*
- ***Underfitting:*** *A small margin yields small gradients for incorrectly ranked samples, stifling learning where it is most needed.*

Table 1 provides a concrete illustration of these cases. Resolving this dilemma requires a margin that adheres to a key principle: it must be **dynamically adaptive** to the policy's instance-wise ranking accuracy. This motivates the concept of an Oracle Ranking Margin.

Definition 2 (Oracle Ranking Margin) *The **Oracle Ranking Margin**, γ^*, is an instance-specific, non-negative threshold that acts as a dynamic learning target [3], distinguishing between incorrectly ranked samples that require a strong learning signal $r_{\pi_\theta}(x, y_w, y_l) < \gamma^*$, and correctly ranked ones whose gradients should be suppressed to prevent overfitting $r_{\pi_\theta}(x, y_w, y_l) > \gamma^*$. The instance-wise ranking accuracy $r_{\pi_\theta}(x, y_w, y_l)$ is defined as $h_w(\log \pi_w) - h_l(\log \pi_l)$.*

With these principles established, to thoroughly investigate whether existing DPO-style algorithms satisfy the above demands, our analysis focuses on two of the most celebrated objectives, including DPO and SimPO.

Analysis of DPO. The margin, $\gamma = \beta \log \pi_{\text{ref}}(y_w|x) - \beta \log \pi_{\text{ref}}(y_l|x)$ derived from the reference model, can neither dynamically adapt to instance-wise ranking correctness nor remaining positive. This irrelevance suggests that DPO can reproduce all cases above, which might result in overfitting and underfitting, as illustrated by case 2 and case 3.

[3]We provide further theoretical analysis in Extended version.

Analysis of SimPO. Although $\gamma = C > 0$ is guaranteed to remain positive, it overlooks the variability inherent in instance-wise ranking correctness. This rigidity suggests that SimPO can reproduce case 2 and case 4, which could lead to overfitting as shown in case 2.

Our analysis reveals that margin designs of both DPO and SimPO fail to proactively adapt based on instance-wise ranking correctness, leading to overfitting and underfitting in preference learning and, consequently, suboptimal ranking accuracy and alignment performance.

4.2 AMaPO: Adaptive Margin-attached Preference Optimization

To overcome the aforementioned limitations, we introduce **Adaptive Margin-attached Preference Optimization (AMaPO)**. Our approach is designed to directly resolve the Overfitting-Underfitting Dilemma by dynamically adjusting the gradient magnitude for each sample. We build upon the robust SimPO (Meng, Xia, and Chen 2024) framework, but our key innovation lies in replacing its static margin with a principled, instance-wise adaptive margin, $\gamma(x, y_w, y_l)$. The general objective is:

$$\begin{aligned}\mathcal{L}_{\text{AMaPO}}(\pi_\theta; \mathcal{D}) = -\,\mathbb{E}_{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) \sim \mathcal{D}} \big[\log \sigma \big(r_{\pi_\theta}(x, y_w, y_l) \\ - \gamma(x, y_w, y_l) \big) \big],\end{aligned} \tag{7}$$

where $r_{\pi_\theta}(x, y_w, y_l) = \frac{\beta}{|y_w|} \log \pi_\theta(y_w|x) - \frac{\beta}{|y_l|} \log \pi_\theta(y_l|x)$ is the implicit ranking accuracy, and $m(x) = \log \sigma(x)$ as Eq. 2. The following subsections detail the design of our adaptive margin, $\gamma(x, y_w, y_l)$.

Formulating the Ideal Adaptive Margin. Our theoretical analysis concluded that an ideal margin must be *dynamically adaptive*. To satisfy this, from definition 2, the margin should reflect the gap between the policy's performance and the ideal target set by the Oracle Ranking Margin, γ^*. We therefore formulate the ideal adaptive margin as:

$$\gamma^*(x, y_w, y_l) = \mathbb{I}\left[(\gamma^* - r_{\pi_\theta}(x, y_w, y_l)) > 0\right] \cdot \gamma^*. \tag{8}$$

This formulation is inherently dynamic and directly addresses the dilemma. For incorrectly ranked samples where $r_{\pi_\theta} \leq 0 < \gamma^*$, the margin is positive and large, amplifying the corrective gradient to mitigate underfitting. For correctly ranked samples where $r_{\pi_\theta} > \gamma^* \geq 0$, the margin becomes zero, suppressing the gradient and preventing overfitting.

Estimating the Oracle Margin in Practice. Since the true Oracle Ranking Margin γ^* is inaccessible, we must estimate it. Motivated by empirical findings that the implicit reward distributions of aligned models are often right-skewed (Qin, Feng, and Yang 2024), we propose using the mean implicit margin μ_r within the current training batch B as a robust, annotation-free proxy for γ^*. To make the ideal margin from Eq. 8 tractable by relaxing the non-negativity constraint (i.e., the $\max(\cdot, 0)$ function), while also stabilizing the estimation and scaling it appropriately, we apply Z-score normalization (Patro and Sahu 2015). This yields:

$$\gamma(x, y_w, y_l) = \max\left(\frac{\mu_r - r_{\pi_\theta}(x, y_w, y_l)}{\sigma_r} \cdot \mu_r, 0\right), \quad (9)$$

where μ_r and σ_r are the mean and standard deviation of $r_{\pi_\theta}(x, y_w, y_l)$ computed within batch B. The term $\frac{\mu_r - r_{\pi_\theta}}{\sigma_r}$ calculates the normalized "difficulty" of the sample, which is then scaled by the estimated oracle target μ_r itself.

Margin Scaling for Quality Representation. The margin from Eq. 9 effectively captures the relative difficulty of a sample. However, empirical results find that the log probability might not truthfully reflect the quality of generated sequence (Holtzman et al. 2021). To better represent the quality gap between responses and speed up the training of the hard incorrectly ranked samples, we introduce a scaling function. Inspired by the strong correlation between perplexity (PPL) and generation quality (Marion et al. 2023; Gonen et al. 2023), we use an exponential scaling function [4]:

$$h_\gamma(\gamma) = \begin{cases} 0 & \text{if } \gamma = 0, \\ \beta \cdot e^\gamma & \text{if } \gamma > 0. \end{cases} \quad (10)$$

Final Objective. Finally, to ensure that our adaptive margin serves as a fixed target for each sample within an optimization step, we apply a stop-gradient operation ($\text{sg}[\cdot]$) to prevent the gradient from margin calculation. Incorporating all components, we obtain the final AMaPO objective:

$$\mathcal{L}_{\text{AMaPO}}(\pi_\theta; \mathcal{D}) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) \sim \mathcal{D}}\big[\log \sigma\big(r_{\pi_\theta}(x, y_w, y_l) - h_\gamma(\text{sg}[\gamma(x, y_w, y_l)])\big)\big]. \quad (11)$$

In summary, by employing an adaptive margin derived from the model's current ranking correctness, AMaPO effectively manipulates the gradient to focus on learning from its errors. This design directly resolves the overfitting and underfitting issues inherent in prior methods, leading to improved ranking accuracy and superior alignment performance.

5 Experiments

This section presents a series of experiments designed to rigorously evaluate AMaPO from multiple perspectives. Our primary goal is to establish its overall superiority against current methods, measured by both ranking accuracy and performance on downstream tasks. Furthermore, we provide a targeted analysis to confirm that AMaPO directly addresses the core overfitting and underfitting issues identified in our framework. Finally, detailed ablation studies to validate the importance of each components.

[4]This elegantly transforms our additive, log-space margin into the geometric mean of the PPL ratio within the batch. The derivation is in Extended version.

5.1 Experimental Setup

Basemodels. To assess the robustness and general applicability of our method, we experiment with two widely-used open-source model families: Llama3-8B (AI@Meta 2024) and Mistral-7B (Jiang 2024). For each family, we test two distinct scenarios: (1) a **Base** model setup, where we first perform SFT before preference alignment, and (2) an **Instruct** model setup, where we begin directly with the publicly available instruction-tuned model.

Training Datasets. Our training pipeline for the **Base** models follows the well-established recipe (Tunstall et al. 2023): we fine-tune on UltraChat-200k (Ding et al. 2023) to create an SFT model, which then is used to for preference optimization on the UltraFeedBack Binarized dataset (Cui et al. 2023). For the **Instruct** setup, to mitigate the distribution shift between the instruction tuned model and the preference data (Meng, Xia, and Chen 2024), we regenerate 5 candidate responses for each prompt from UltraFeedback by the Instruct model. We then use PairRM (Jiang, Ren, and Lin 2023) to score the 5 responses to construct the chosen and rejected response pairs.

Evaluation benchmarks. Our evaluation is designed to be multi-faceted, assessing performance from ranking accuracy to its practical downstream impact. (1) **Ranking Accuracy:** To directly validate our claims about improving preference learning, we measure ranking accuracy on RM-Bench (Liu et al. 2024), a challenging benchmark designed to test a model's grasp of subtle preference nuances. (2) **Overfitting-underfitting Problem:** To further test our hypothesis about resolving this problem, we also evaluate ranking accuracy across four generalization scenarios on UltraFeedBack following (Hong et al. 2025): In-Distribution (ID), Prompt-OOD, Response-OOD, and Mutual-OOD. (3) **Downstream Performance:** To confirm that improved ranking translates to better downstream performance, we evaluate on the popular AlpacaEval 2 (Li et al. 2023) and MT-Bench (Zheng et al. 2023a) benchmarks.

Baselines. We compare AMaPO against other advanced offline preference optimization methods, including DPO (Rafailov et al. 2023), SLiC (Zhao et al. 2023), IPO (Azar et al. 2024), KTO (Ethayarajh et al. 2024), and SimPO (Meng, Xia, and Chen 2024). To ensure a fair and rigorous comparison, we have thoroughly tuned the hyperparameters for each baseline method and report their best performance.

5.2 Main Results on Benchmarks

Results on Preference Ranking and Downstream Benchmarks. Our main results demonstrate that AMaPO achieves superior performance in both preference learning and downstream alignment tasks. As shown in Table 3, which evaluates the ability to distinguish subtle preference differences, AMaPO consistently outperforms other methods in different scenarios. Specifically, on the Llama3-8B-Base setup, AMaPO improves the ranking accuracy over the strong SimPO baseline by 2.4 and 2.1 points on Normal and Hard cases, respectively. This highlights its enhanced capability to discern subtle quality variations. In contrast,

Method	Llama3-8B-Base			Llama3-8B-Instruct			Mistral-7B-Base			Mistral-7B-Instruct		
	AlpacaEval 2		MT-Bench	AlpacaEval 2		MT-Bench	AlpacaEval 2		MT-Bench	AlpacaEval 2		MT-Bench
	LC (%)	WR (%)	GPT-4 Turbo	LC (%)	WR (%)	GPT-4 Turbo	LC (%)	WR (%)	GPT-4 Turbo	LC (%)	WR (%)	GPT-4 Turbo
SFT	6.2	4.6	3.3	26.0	25.3	6.9	8.4	6.2	4.8	17.1	14.7	6.2
DPO	18.2	15.5	6.5	40.3	37.9	7.0	15.1	12.5	5.9	26.8	24.9	6.3
SLiC	12.3	13.7	6.3	26.9	27.5	6.8	10.9	8.9	5.8	24.1	24.6	6.5
IPO	14.4	14.2	6.5	35.6	35.6	7.0	11.8	9.4	5.5	20.3	20.3	6.4
KTO	14.2	12.4	6.3	33.1	31.8	6.9	13.1	9.1	5.4	24.5	23.6	6.4
CPO	10.8	8.1	6.0	28.9	32.2	7.0	9.8	8.9	5.4	23.8	28.8	6.3
SimPO	22.0	20.3	6.6	44.7	40.5	7.0	21.5	**20.8**	6.0	32.1	34.8	6.6
α-DPO	21.7	20.6	**6.8**	**46.6**	39.6	**7.2**	17.2	13.0	**6.2**	34.2	33.8	**6.7**
AMaPO	**26.4**	**21.4**	6.4	46.1	**41.3**	**7.2**	**24.3**	20.6	**6.2**	**34.5**	**35.1**	**6.7**

Table 2: AlpacaEval2 (Li et al. 2023) and MT-Bench (Zheng et al. 2023a) results under the four settings. LC and WR denote length-controlled and raw win rate, respectively. The best results are highlighted in bold.

while DPO often achieves the highest accuracy on Easy cases where style biases can guide the model, it exhibits a significant performance drop on Normal and Hard ones. For instance, on Mistral-7B-Base model, DPO's accuracy plummets from 89.8% on Easy cases to just 18.7% on Hard ones.

This discrepancy aligns perfectly with our gradient analysis, which identifies DPO's tendency to overfit to simple patterns while underfitting to complex, misranked samples. Although SFT models sometimes perform well, especially on instruct-tuned setups, this strength does not reliably carry over to downstream tasks. In contrast, AMaPO's robust preference learning translates directly into superior alignment. As shown in Table 2, AMaPO consistently leads on AlpacaEval 2 and MT-Bench. Notably, it achieves a length-controlled (LC) win rate up to 4.4 points higher than SimPO on Llama3-8B-Base and demonstrates strong performance gains regardless of the evaluation metric. These findings confirm that AMaPO effectively enhances core ranking capabilities, and this fundamental improvement successfully generalizes to diverse downstream applications. [5]

	Method	Avg.	Easy	Normal	Hard
Mistral-7B Base	DPO	55.8	**89.8**	58.7	18.7
	SimPO	56.5	86.3	58.7	23.5
	α-DPO	**58.4**	86.6	**62.7**	25.0
	AMaPO	58.1	86.4	62.5	**25.4**
Llama3-8B Base	DPO	54.6	**91.6**	57.4	15.0
	SimPO	56.9	87.5	60.2	23.3
	α-DPO	58.2	87.1	62.1	**25.6**
	AMaPO	**58.6**	87.8	**62.6**	25.4
Mistral-7B Instruct	DPO	53.2	91.3	55.5	12.9
	SimPO	54.9	89.0	57.8	**17.8**
	α-DPO	55.2	90.9	58.4	16.3
	AMaPO	**55.5**	**91.4**	**59.2**	15.9
Llama3-8B Instruct	DPO	53.4	**89.6**	55.4	15.1
	SimPO	55.7	86.1	58.7	22.4
	α-DPO	55.0	83.4	57.6	23.8
	AMaPO	**56.5**	85.7	**59.8**	**24.0**

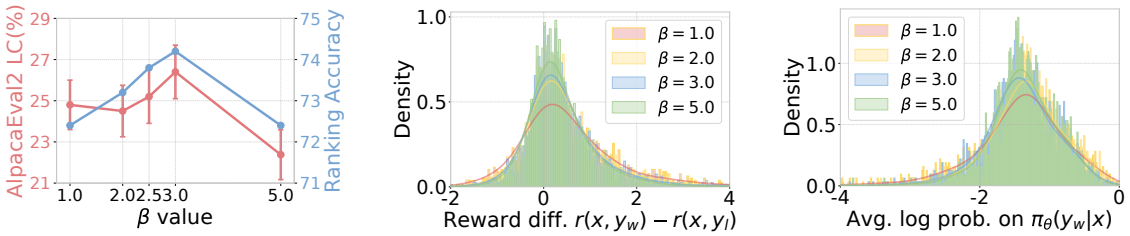
Table 3: RM-bench (Liu et al. 2024) results under the three setups, where Easy, Normal and Hard represent the preferred responses have the better, same and worse style compared to less preferred ones. Our AMaPO can achieve good performance on distinguishing responses across various setups.

	Method	$\mathcal{D}_{\text{ID}}$	$\mathcal{D}_{\sim\text{Prompt}}$	$\mathcal{D}_{\sim\text{Response}}$	$\mathcal{D}_{\sim\text{Mutual}}$
Mistral-7B Base	DPO	69.66	83.01	68.41	67.35
	SimPO	78.02	91.05	76.08	75.94
	AMaPO	**79.26**	**92.46**	**83.62**	**82.97**
Llama3-8B Base	DPO	64.73	76.2	61.29	60.25
	SimPO	**78.1**	91.14	**78.6**	76.3
	AMaPO	77.33	**91.41**	77.87	**76.91**
Mistral-7B Instruct	DPO	82.19	86.83	75.67	72.89
	SimPO	81.39	**95.17**	86.31	85.64
	AMaPO	**82.85**	94.88	**91.74**	**90.91**
Llama3-8B Instruct	DPO	79.23	87.34	67.94	66.16
	SimPO	**84.45**	95.68	93.46	93.16
	AMaPO	84.04	**96.5**	**94.16**	**94.26**

Table 4: Overfitting-underfitting Problem over ultrafeedback dataset following (Lin et al. 2024; Hong et al. 2025). Our AMaPO achieves promising ranking accuracy across in-domain and out-of-distribution settings.

Results on the Overfitting-Underfitting Problem. To further validate our theoretical claims regarding the overfitting-underfitting dilemma, we evaluate model performance across four generalization scenarios. As shown in Table 4, the results strongly support our analysis and demonstrate the effectiveness of AMaPO. DPO consistently performs the worst across both in-distribution and OOD settings, confirming its dual vulnerability. For example, on the Llama3-8B-Base setup, DPO's accuracy is not only the lowest in-distribution but also fails to generalize, lagging behind AMaPO by over 15 points on Prompt OOD. This aligns with our theory that DPO underfits on challenging OOD samples. SimPO, while achieving high in-distribution accuracy, suggesting a tendency to overfit to seen data, does not generalize as effectively as AMaPO due to its static margin. For instance, on Mistral-7B-Base, SimPO's performance drops by 15 points from Prompt OOD to Mutual OOD. In contrast, AMaPO resolves this tension: it matches or exceeds

[5]For more results like further downstream task results, generation length, batch size ablation, and cases, see Extended version.



(a) Performance w/ different β (b) Prob. diff. distribution. (c) Log prob. distribution.

Figure 1: Ablation studies of the β. (a) Ranking accuracy and AlpacaEval2 LC win rate under different β values. (b) Reward difference distribution under different β values. (c) Log likelihood distribution on chosen responses under different β values.

SimPO's in-distribution performance while achieving the best results across all OOD settings. This robust generalization confirms that our adaptive margin successfully mitigates the overfitting-underfitting problem we identified.

5.3 Ablations and Further Analysis

This section delves deeper into the components and behavior of AMaPO. We begin by ablating the modules, such as Z-normalization and exponential scaling, to confirm their necessity. Our main analysis then focuses on the hyperparameter β, which controls the strength of the adaptive margin.

Importance of the components of AMaPO. Due to the significant computational resources required, we conduct this ablation study on Llama3-8B models, evaluating performance on key downstream benchmarks. Results in Table 5 show that the complete AMaPO design, which integrates Z-normalization (can also been shown as an alternative of linear scaling functions), exponential scaling, and a zero margin for correctly ranked samples, is critical for achieving robust and well-rounded performance. While most ablated variants still outperform the DPO baseline, removing any single component results in a notable performance degradation. This is particularly evident on the Llama3-8B-Base, where removing exponential scaling and Z-normalization ('w/o adaptive') degrades the AlpacaEval 2 LC and WR to 20.7% and 16.7%, below the full AMaPO's 26.3% and 21.4%. Similarly, removing any components of AMaPO on the Llama3-8B-Instruct causes the score on MT-Bench to drop from AMaPO's 7.2 to about 7.0. Thus, the synergistic combination of all components in AMaPO is essential for a better adaptive margin and ensuring consistently superior performance across diverse evaluation settings.

Method	Llama3-8B-Base			Llama3-8B-Instruct		
	AlpacaEval 2		MT-Bench	AlpacaEval 2		MT-Bench
	LC (%)	WR (%)	GPT-4 Turbo	LC (%)	WR (%)	GPT-4 Turbo
DPO	18.2	15.5	6.5	40.3	37.9	7.0
SimPO	22.0	20.3	**6.6**	44.7	40.5	7.0
AMaPO	**26.3**	**21.4**	6.5	46.1	41.3	**7.2**
w/o Z-norm	24.8	20.4	6.3	44.2	39.6	7.0
w/o exp	24.0	17.6	6.1	47.1	42.9	7.0
w/o adaptive	20.7	16.4	6.3	47.2	42.5	7.0
w/o zero	22.3	21.1	**6.6**	**48.6**	**45.0**	6.7

Table 5: Ablation studies under Llama3-8B setups. We ablate the key design of AMaPO, the adaptive margin.

Influence of β. To investigate the influence of our sole hyperparameter β, we trained AMaPO on the UltraFeedback dataset using the Llama3-8B-Base model with varying β values. As shown in Figure 1a, our analysis on the AlpacaEval 2 and UltraFeedback test sets reveals a clear inverted U-shaped relationship between β and performance. Both ranking accuracy and the LC win rate initially increase with β before declining, indicating that an optimal regularization strength is empirically around $\beta = 3$. The underlying reason for this trend is revealed by analyzing the probability distributions on the UltraFeedback test set (Figures 1b and 1c). As β increases, the distribution of the probability margin, $\pi_\theta(y_w|x) - \pi_\theta(y_l|x)$, sharpens significantly, but the value at the peak of the likelihood of winning responses, $\pi_\theta(y_w|x)$, decreases. This reveals a critical trade-off: while a stronger regularization (a higher β) initially forces the model to more accurately fit our adaptive margin, leading to performance gains, an excessively high β causes the likelihood distribution to become overly peaked, resulting in model degeneration and a subsequent drop in performance.

6 Conclusion

In this work, we present a novel analysis of offline preference optimization and propose a new algorithm, AMaPO, to address the dilemma we identified. Through a unified margin-based framework, our gradient analysis reveals that DPO-style methods suffer from an overfitting-underfitting dilemma, inefficiently learning from preference data. AMaPO rectifies this by introducing an instance-wise adaptive margin that intelligently prioritizes difficult, misranked samples while ignoring those already learned. Extensive experiments show that AMaPO has a better ranking accuracy, leading to a better performance on downstream tasks. By providing both a more effective alignment algorithm and a sharper analytical lens, this work paves the way for more principled advances in preference optimization. [6]

[6] A detailed discussion of limitations is in Extended version.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62272330, and No. U24A20328), and in part by the Science Fund for Creative Research Groups of Sichuan Province Natural Science Foundation (No. 2024NSFTD0035).

References

AI@Meta. 2024. Llama 3 Model Card.

Amini, A.; Vieira, T.; and Cotterell, R. 2024. Direct Preference Optimization with an Offset. In *ACL (Findings)*.

Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 4447–4455. PMLR.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Boser, B. E.; Guyon, I. M.; and Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Chen, A.; Malladi, S.; Zhang, L.; Chen, X.; Zhang, Q. R.; Ranganath, R.; and Cho, K. 2024. Preference learning algorithms do not learn preference rankings. *Advances in Neural Information Processing Systems*, 37: 101928–101968.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; Zhu, W.; Ni, Y.; Xie, G.; Liu, Z.; and Sun, M. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3029–3051.

Engstrom, L.; Ilyas, A.; Santurkar, S.; Tsipras, D.; Janoos, F.; Rudolph, L.; and Madry, A. 2020. Implementation matters in deep policy gradients: A case study on ppo and trpo. *arXiv preprint arXiv:2005.12729*.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. arXiv:2402.01306.

Feng, D.; Qin, B.; Huang, C.; Zhang, Z.; and Lei, W. 2024. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*.

Gonen, H.; Iyer, S.; Blevins, T.; Smith, N. A.; and Zettlemoyer, L. 2023. Demystifying Prompts in Language Models via Perplexity Estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10136–10148.

Gorbatovski, A.; Shaposhnikov, B.; Malakhov, A.; Surnachev, N.; Aksenov, Y.; Maksimov, I.; Balagansky, N.; and Gavrilov, D. 2024. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*.

Holtzman, A.; West, P.; Shwartz, V.; Choi, Y.; and Zettlemoyer, L. 2021. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7038–7051.

Hong, J.; Lee, N.; Kim, E.; Son, G.; Chung, W.; Gupta, A.; Tang, S.; and Thorne, J. 2025. On the Robustness of Reward Models for Language Model Alignment. *arXiv preprint arXiv:2505.07271*.

Jiang, D.; Ren, X.; and Lin, B. Y. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14165–14178.

Jiang, F. 2024. *Identifying and mitigating vulnerabilities in llm-integrated applications*. Master's thesis, University of Washington.

Kim, K.; Seo, A.; Liu, H.; Shin, J.; and Lee, K. 2024. Margin Matching Preference Optimization: Enhanced Model Alignment with Granular Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 13554–13570.

Li, X.; Zhang, T.; Dubois, Y.; Taori, R.; Gulrajani, I.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval.

Lin, Y.; Seto, S.; Ter Hoeve, M.; Metcalf, K.; Theobald, B.-J.; Wang, X.; Zhang, Y.; Huang, C.; and Zhang, T. 2024. On the limited generalization capability of the implicit reward model induced by direct preference optimization. *arXiv preprint arXiv:2409.03650*.

Liu, T.; Yu, X.; Zhou, W.; Gu, J.; and Tresp, V. 2025. FocalPO: Enhancing Preference Optimizing by Focusing on Correct Preference Rankings. *arXiv preprint arXiv:2501.06645*.

Liu, Y.; Liu, P.; and Cohan, A. 2024. Understanding reference policies in direct preference optimization. *arXiv preprint arXiv:2407.13709*.

Liu, Y.; Yao, Z.; Min, R.; Cao, Y.; Hou, L.; and Li, J. 2024. RM-bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*.

Ma, Q.; Shi, J.; Jin, C.; Hwang, J.-N.; Belongie, S.; and Li, L. 2025. Gradient Imbalance in Direct Preference Optimization. *arXiv preprint arXiv:2502.20847*.

Marion, M.; Üstün, A.; Pozzobon, L.; Wang, A.; Fadaee, M.; and Hooker, S. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Patro, S.; and Sahu, K. K. 2015. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Qin, B.; Feng, D.; and Yang, X. 2024. Towards understanding the influence of reward margin on preference model performance. *arXiv preprint arXiv:2404.04932*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Tajwar, F.; Singh, A.; Sharma, A.; Rafailov, R.; Schneider, J.; Xie, T.; Ermon, S.; Finn, C.; and Kumar, A. 2024. Preference Fine-Tuning of LLMs Should Leverage Suboptimal, On-Policy Data. In *International Conference on Machine Learning*, 47441–47474. PMLR.

Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; Von Werra, L.; Fourrier, C.; Habib, N.; et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Turner, H.; and Firth, D. 2012. Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software*, 48: 1–21.

Wu, J.; Wang, X.; Yang, Z.; Wu, J.; Gao, J.; Ding, B.; Wang, X.; and He, X. 2024. $\backslash\alpha$-DPO: Adaptive Reward Margin is What Direct Preference Optimization Needs. *arXiv preprint arXiv:2410.10148*.

Xiao, T.; Yuan, Y.; Chen, Z.; Li, M.; Liang, S.; Ren, Z.; and Honavar, V. G. 2025. SimPER: A Minimalist Approach to Preference Alignment without Hyperparameters. *arXiv preprint arXiv:2502.00883*.

Xiao, T.; Yuan, Y.; Zhu, H.; Li, M.; and Honavar, V. G. 2024. Cal-dpo: Calibrated direct preference optimization for language model alignment. *arXiv preprint arXiv:2412.14516*.

Yan, Y.; Miao, Y.; Li, J.; Zhang, Y.; Xie, J.; Deng, Z.; and Yan, D. 2024. 3d-properties: Identifying challenges in dpo and charting a path forward. *arXiv preprint arXiv:2406.07327*.

Yuan, H.; Zeng, Y.; Wu, Y.; Wang, H.; Wang, M.; and Leqi, L. 2024. A Common Pitfall of Margin-based Language Model Alignment: Gradient Entanglement. *arXiv preprint arXiv:2410.13828*.

Zhao, H.; Winata, G. I.; Das, A.; Zhang, S.-X.; Yao, D. D.; Tang, W.; and Sahu, S. 2024. Rainbowpo: A unified framework for combining improvements in preference optimization. *arXiv preprint arXiv:2410.04203*.

Zhao, Y.; Joshi, R.; Liu, T.; Khalman, M.; Saleh, M.; and Liu, P. J. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zheng, R.; Dou, S.; Gao, S.; Hua, Y.; Shen, W.; Wang, B.; Liu, Y.; Jin, S.; Liu, Q.; Zhou, Y.; et al. 2023b. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.